METHOD AND DEVICE FOR SIMULTANEOUS VOICE RECOGNITION, SPEAKER SEGMENTATION AND SPEAKER CLASSIFICATION

Patent Number: JP2001060098
Publication date: 2001-03-06
Inventor(s): HAMEION SADARU MOHAMMAD BEIGI; TRITSCHLER ALAIN CHARLES LOUIS; VISWANATHAN MAHESH
Applicant(s): INTERNATL BUSINESS MACH CORP
Requested Patent: JP2001060098
Application Number: JP20000188625 20000623
Priority Number (s):
IPC Classification: G10L15/04; G06F3/16; G10L17/00; G10L15/00; G10L15/28
EC Classification:
Equivalents: CN1279462

Abstract

PROBLEM TO BE SOLVED: To obtain a method, in which audio information from an audio/video source is automatically transferred and a speaker is identified simultaneously, by tranferring the audio source, simultaneously identifying latent segment boundaries and assigning a speaker label to each identified segment.
SOLUTION: The method includes a step, in which a transfer is made for an audio source to generate a text version of audio information, a step which simultaneously identifies latent segment boundaries, and a step in which a speaker label is assigned to each of identified segments. A simultaneous transfer, segmentation and speaker identification process 500 generates a transfer of audio information, which represents a speaker related to each segment, in real time. A segmentation process 600 identifies all frames in which segment boundaries may exist. A speaker identifying process 700 assigns a speaker label to each of the segments that use registered speaker databases.

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2001－60098
（P2001－60098A）
(43)公開日　平成13年３月６日(2001.3.6)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　15/04		Ｇ１０Ｌ　3/00		５１５Ｃ
Ｇ０６Ｆ　3/16	３２０	Ｇ０６Ｆ　3/16		３２０Ｇ
Ｇ１０Ｌ　17/00		Ｇ１０Ｌ　3/00		５４５Ａ
15/00				５５１Ｐ
15/28				

審査請求　有　　請求項の数23　ＯＬ　（全 15 頁）

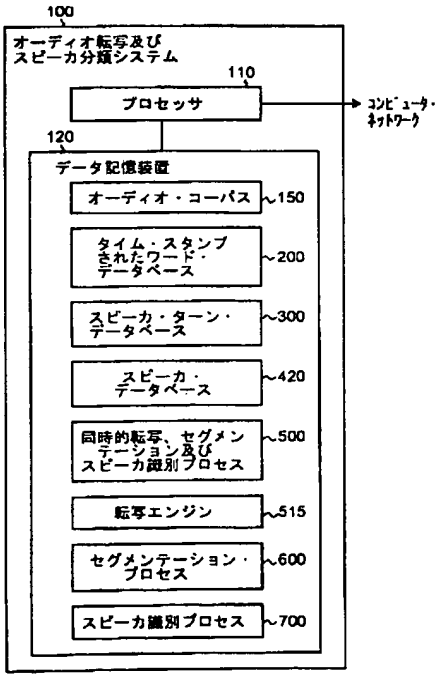
(21)出願番号　特願2000－188625(P2000－188625)
(22)出願日　平成12年６月23日(2000.6.23)
(31)優先権主張番号　０９／３４５２３７
(32)優先日　平成11年６月30日(1999.6.30)
(33)優先権主張国　米国（ＵＳ）

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）
(72)発明者　ハメイオン・サダル・モハマド・ベイギ
アメリカ合衆国ニューヨーク州、ヨークタウン・ハイツ、エッジヒル・ロード　3616
(74)代理人　100086243
弁理士　坂口　博　（外２名）

最終頁に続く

(54)【発明の名称】　同時的な音声認識、スピーカ・セグメンテーション及びスピーカ分類のための方法及び装置

(57)【要約】　（修正有）
【課題】オーディオ／ビデオ・ソースからオーディオ情報を自動的に転写し、同時にスピーカを識別する方法及び装置。
【解決手段】オーディオ転写及びスピーカ分類システムは音声認識システム、スピーカ・セグメンテーション・システム及びスピーカ識別システムを含む。音声認識システムは、各ワードに対してタイム・アライメントを伴う転写物を作成する。スピーカ・セグメンテーション・システムはスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在する可能性のあるすべてのフレームを識別する。スピーカ識別システムは、各識別されたセグメントにスピーカを割当てるため、登録済みのスピーカ・データベースを使用する。オーディオ／ビデオ・ソースからのオーディオ情報は、セグメント境界を識別するため同時に転写及びセグメント化された後、スピーカ識別システムは識別されたテキストの各部分にスピーカ・ラベルを割当てる。



1

【特許請求の範囲】

【請求項１】１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、

前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、

前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、

識別された各セグメントにスピーカ・ラベルを割り当てるステップと、を含む方法。

【請求項２】前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写を作成することを特徴とする請求項１に記載の方法。

【請求項３】前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項１に記載の方法。

【請求項４】前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項１に記載の方法。

【請求項５】前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項１に記載の方法。

【請求項６】前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも１つの別の選択を割り当てるステップを更に含むことを特徴とする請求項１に記載の方法。

【請求項７】前記転写するステップ、識別するステップ、及び割り当てるステップはマルチ・スレッド環境では並列処理ブランチにおいて行われることを特徴とする請求項１に記載の方法。

【請求項８】前記識別するステップはＢＩＣモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項１に記載の方法。

【請求項９】１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、

前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、

（ａ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、

（ｂ）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び

（ｃ）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するステップと、を含む方法。

【請求項１０】前記フィーチャ・ベクトルは共用メモリ・アーキテクチャを使用して前記並列処理ブランチに適用されることを特徴とする請求項９に記載の方法。

2

【請求項１１】前記共用メモリ・アーキテクチャは前記計算されたフィーチャ・ベクトルを前記並列処理ブランチの各々に対応するチャネルに分配することを特徴とする請求項１０に記載の方法。

【請求項１２】前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写物を作成することを特徴とする請求項９に記載の方法。

【請求項１３】前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項９に記載の方法。

【請求項１４】前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項９に記載の方法。

【請求項１５】前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項９に記載の方法。

【請求項１６】前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも１つの別の選択を割り当てるステップを更に含むことを特徴とする請求項９に記載の方法。

【請求項１７】前記識別するステップはＢＩＣモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項９に記載の方法。

【請求項１８】１つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、

コンピュータ読み取り可能なコードを記憶するメモリと、

前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、

を含み、

前記コンピュータ読み取り可能なコードは、

前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するように、

前記転写と実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するように、及び各識別されたセグメントにスピーカ・ラベルを割り当てるように、

構成されることを特徴とするシステム。

【請求項１９】コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、

前記コンピュータ読み取り可能なプログラム・コード手段は、

オーディオ情報のテキスト・バージョンを作成するため

にオーディオ・ソースを転写するステップと、
前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、
識別された各セグメントにスピーカ・ラベルを割り当てるステップと、
を含むことを特徴とする製造物。
【請求項２０】１つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、
コンピュータ読み取り可能なコードを記憶するメモリと、
前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、
を含み、
前記コンピュータ読み取り可能なコードは、
前記オーディオ情報からフィーチャ・ベクトルを計算し、
（ｉ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、
（ii）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び
（iii）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するように構成されることを特徴とするシステム。
【請求項２１】コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、
前記コンピュータ読み取り可能なプログラム・コード手段は、
前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、
（ｉ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、
（ii）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び
（iii）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するステップと、を含むことを特徴とする製造物。
【請求項２２】１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、
前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、
前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、
識別された各セグメントにスピーカ・ラベルを割り当てるステップと、
前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、
を含む方法。
【請求項２３】１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、
前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、
（ｉ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、
（ii）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び
（iii）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するステップと、
前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、
を含む方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、概して云えば、オーディオ情報分類システムに関し、詳しく云えば、オーディオ情報を転写（ｔｒａｎｓｃｒｉｂｅ）し、オーディオ・ファイルにおけるスピーカ（発声者）を識別するための方法及び装置に関するものである。
【０００２】
【従来の技術】放送ニュース機構及び情報検索サービスのような多くの機構は、記憶及び検索のために大量のオーディオ情報を処理しなければならない。オーディオ情報は、主題又はスピーカの名前、或いはそれらの両方によって分類されなければならないことが多い。主題によってオーディオ情報を分類するためには、先ず、音声認識システムが、自動分類又はインデキシングのために、オーディオ情報をテキストの形に転写する。しかる後、照会／ドキュメント・マッチングを行って関連ドキュメントをユーザに戻すためにインデックスが使用可能である。
【０００３】従って、主題によってオーディオ情報を分類するというプロセスは本質的には完全に自動化されたものになっている。しかし、スピーカによってオーディオ情報を分類するというプロセスは、特に、放送ニュースのようなリアルタイムの応用に対しては、大きな労力を要する仕事を残すことが多い。スピーカ登録情報を使用してオーディオ・ソースからスピーカを自動的に識別するための数多くの計算主体のオフライン・テクニックが提案されているけれども、スピーカ分類プロセスはヒューマン・オペレータによって最も頻繁に行われ、ヒューマン・オペレータは各スピーカの変更を識別し、対応

するスピーカの識別を行う。

【０００４】本発明の親出願（１９９９年４月９日出願の米国特許出願番号０９/２８８,７２４号）は、オーディオ・コンテント（主題）及びスピーカのアイデンティティに基づいてオーディオ情報を検索するための方法及び装置を開示している。タイム・スタンプされたコンテント・インデックス・ファイル及びスピーカ・インデックス・ファイルを作成するために、インデキシング・システムがオーディオ情報を転写し、そしてインデックスする。しかる後、その生成されたコンテント及びスピーカ・インデックスは、オーディオ・コンテント及びスピーカ・アイデンティティに基づいて照会／ドキュメント・マッチングを行うために利用可能である。オーディオ・ソースからオーディオ情報を自動的に転写し、同時にスピーカをリアルタイムで識別する方法及び装置に対する要求が存在する。ベイズ情報基準（Baysian Information Criterion－ＢＩＣ）に基づいてスピーカ・セグメンテーション及びクラスタリングの改良を提供する方法及び装置に対する更なる要求も存在する。

【０００５】

【発明が解決しようとする課題】従って、本発明の目的は、オーディオ／ビデオ・ソースからのオーディオ情報を自動的に転写し、同時にスピーカを識別するための方法及び装置を開示することにある。

【０００６】

【課題を解決するための手段】開示されたオーディオ転写及びスピーカ分類システムは、音声認識システム、スピーカ・セグメンテーション・システム、及びスピーカ識別システムを含む。本発明の１つの局面によれば、オーディオ情報は、マルチスレッド環境における並列ブランチに沿って音声認識システム、スピーカ・セグメンテーション・システム、及びスピーカ識別システムによって処理される。

【０００７】音声認識システムは、転写物を、その転写物内の各ワードに対するタイム・アライメントを伴って作成する。スピーカ・セグメンテーション・システムはスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別する。しかる後、スピーカ識別システムは、登録されたスピーカ・データベースを使用して、各識別されたセグメントにスピーカを割り当てる。

【０００８】本発明は、マルチスレッド環境における並列ブランチに沿って音声認識システム、スピーカ・セグメンテーション・システム、及びスピーカ識別システムによって処理されるフィーチャ・ベクトルを計算するために共通のフロント・エンド処理を利用する。一般に、フィーチャ・ベクトルは、例えば、計算されたフィーチャ・ベクトルを各チャネル（各処理スレッドに対応する）に分配するするためにサーバのような態様で作用する共用メモリ・アーキテクチャを使用して、３つの多重処理スレッドに分配可能である。

【０００９】本発明のもう１つの局面によれば、オーディオ／ビデオ・ソースからのオーディオ情報が同時に転写され及びセグメント境界を識別するためにセグメント化される。一旦音声セグメントがセグメンテーション・システムによって識別されると、スピーカ識別システムがその転写されたテキストの各部分にスピーカ・ラベルを割り当てる。

【００１０】本願で開示されるセグメンテーション・プロセスは、オーディオ・データを通るパス上に、しかも、転写エンジンと同じパス上にあるセグメント境界であってスピーカ変更に対応するセグメント境界が存在するすべての可能なフレームを識別する。フレームは所定の期間にわたって音声特性を表す。セグメンテーション・プロセスは、２つのモデルを比較するモデル選択基準を使用して、所定のフレームにセグメント境界が存在するかどうかを決定する。第１モデルは、単一の全共分散ガウス分布（ｆｕｌｌ－ｃｏｖａｒｉａｎｃｅ Ｇａｕｓｓｉａｎ）を使用するサンプル（$x_1,\ldots,x_n$）のウインドウ内にセグメント境界が存在しないものと仮定する。第２モデルは、第１ガウス分布から得られた（$x_1,\ldots,x_i$）及び第２ガウス分布から得られた（$x_{i+1},\ldots,x_n$）を持った２つの全共分散ガウス分布を使用するサンプル（$x_1,\ldots,x_n$）のウインドウ内にセグメント境界が存在するものと仮定する。

【００１１】本願で開示されるスピーカ識別システムは、登録されたスピーカ・データベースを使用して各識別されたセグメントにスピーカ・ラベルを割り当てる。スピーカ識別プロセスはセグメンテーション・プロセスによって識別されたターンを、共用のフロント・エンドによって生成されたフィーチャ・ベクトルと共に受け取る。一般に、スピーカ識別システムは登録されたスピーカ・データベースにセグメント発声音（ｕｔｔｅｒａｎｃｅ）を比較し、「最も近似した」スピーカを見つける。そのスピーカ識別システムのためのモデル・ベース方式及びフレーム・ベース方式が開示される。

【００１２】本発明の結果は、例えば、転写されたテキストを、割り当てられたスピーカ・ラベルと共に各セグメントに供給するユーザに直接に出力可能である。更に、本発明の結果は１つ又は複数のデータベースに記録可能であり、オーディオ・コンテント及びスピーカのアイデンティティに基づいてオーディ情報に（及び間接的にはビデオに）参照を行うためにコンテント及びスピーカ・サーチ方法の結果を結合するという前記親出願において開示されたオーディオ検索システムのようなオーディオ検索システムによって利用可能である。

【００１３】以下の詳細な説明及び図面を参照することによって、本発明の更に完全な理解及び本発明の更なる特徴及び利点の理解が得られるであろう。

【００１４】

【発明の実施の形態】図１は、オーディオ／ビデオ・ソースからのオーディオ情報を自動的に転写し、同時にスピーカを識別するという本発明によるオーディオ転写及びスピーカ分類システム１００を示す。オーディオ／ビデオ・ソース・ファイルは、例えば、オーディオ・レコーディングであってもよく、或いは、例えば、放送のニュース・プログラムからの生番組であってもよい。オーディオ／ビデオ・ソースは、先ず、転写され、同時に、スピーカの変更を表すセグメント境界が存在するすべての可能なフレームを識別するように処理される。

【００１５】オーディオ転写及びスピーカ分類システム１００は、音声認識システム、スピーカ・セグメンテーション・システム及びスピーカ識別システムを含む。音声認識システムは、転写物を、その転写物における各ワードに対するタイム・アライメントを伴って作成する。スピーカ・セグメンテーション・システムはスピーカを個別化し、セグメント境界が存在するすべての可能なフレームを識別する。セグメントは、所定のスピーカと関連したオーディオ・ソースの連続部分である。しかる後、スピーカ識別システムが各セグメントにスピーカ・ラベルを割り当てる。

【００１６】図１は、本発明による例示的なオーディオ転写及びスピーカ分類システム１００のアーキテクチャを示すブロック図である。オーディオ転写及びスピーカ分類システム１００は、図１に示された汎用コンピュータ・システムのような汎用コンピュータ・システムとして具体化可能である。そのオーディオ転写及びスピーカ分類システム１００はプロセッサ１１０及びデータ記憶装置１２０のような関連メモリを含む。なお、データ記憶装置１２０は分散型又はローカル型のものでよい。プロセッサ１１０は単一のプロセッサとして、又は並行して動作する複数のローカル・プロセッサ又は分散プロセッサとして実施可能である。データ記憶装置１２０及び／又は読取り専用メモリ（ＲＯＭ）は１つ又は複数の命令を記憶するように動作可能であり、プロセッサ１１０はそれらの命令を検索、解釈、及び実行するように動作可能である。

【００１７】望ましくは、データ記憶装置１２０は、本発明に従ってリアルタイムで処理可能な１つ又は複数の事前記録された又は生のオーディオ・ファイル又はビデオ・ファイル（或いは、それらの両方）を記憶するためのオーディオ・コーパス（ｃｏｒｐｕｓ）データベース１５０を含む。又、データ記憶装置１２０は、図２に関連して後述するタイム・スタンプ・ワード・データベース２００も含む。そのデータベース２００は音声認識システムによって生成されたものであり、一組のタイム・スタンプされたワードを含む。図３に関連して後述するスピーカ・ターン・データベース３００はスピーカ・セグメンテーション・システムと関連してスピーカ識別システムによって作成され、各セグメントの開始時間を、１つ又は複数の対応する提案されたスピーカ・ラベルと共に表示する。図４と関連して後述するスピーカ・データベース４２０はスピーカ登録プロセス４１０によって作成され、各登録されたスピーカに対するエントリを含む。図１の例示的な実施例に示されたそれらの生成されたデータベース２００及び３００は、本発明の結果がリアルタイムでユーザに表示されるオンライン・インプリメンテーションに対しては必要とされず、その後のアクセスのためにも必要とされないことに注意してほしい。

【００１８】更に、図５及び図６に関連して後述するように、データ記憶装置１２０は、同時転写、セグメンテーション及びスピーカ識別プロセス５００、転写エンジン５１５，セグメンテーション・プロセス６００，及びスピーカ識別プロセス７００を含む。同時転写、セグメンテーション及びスピーカ識別プロセス５００は転写エンジン５１５，セグメンテーション・プロセス６００、及びスピーカ識別プロセス７００の実行を調整する。同時転写、セグメンテーション及びスピーカ識別プロセス５００はオーディオ・コーパス・データベース１５０における１つ又は複数のオーディオ・ファイルを分析し、各セグメントと関連するスピーカを表すオーディオ情報の転写をリアルタイムで作成する。セグメンテーション・プロセス６００はスピーカを個別化し、セグメント境界が存在するすべての可能なフレームを識別する。スピーカ識別プロセス７００は、登録されたスピーカ・データベースを使用する各セグメントにスピーカ・ラベルを割り当てる。

【００１９】図２は、音声認識システムによって生成され、一組のタイム・スタンプされたワードを含む例示的なタイム・スタンプ・ワード・データベース２００を示す。そのタイム・スタンプ・ワード・データベース２００は、各々がその実施例における異なるワードと関連するレコード２１１乃至２１４のような複数のレコードを維持する。ワード・ストリング・フィールド２２０において識別された各ワードに対して、タイム・スタンプ・ワード・データベース２００は開始時間フィールド２３０においてそのワードの開始時間を表示する。

【００２０】図３は例示的なスピーカ・ターン・データベース３００を示す。そのデータベース３００は、スピーカ・セグメンテーション・システムと関連してスピーカ識別システムによって作成され、１つ又は複数の対応する提案されたスピーカ・ラベルと共に各セグメントの開始時間を表す。スピーカ・ターン・データベース３００は、各々が実施例における種々のセグメントによって識別されるレコード３０５乃至３０８のような複数のレコードを維持する。フィールド３２０におけるセグメント番号によって識別された各セグメントに対して、スピーカ・ターン・データベース３００は、オーディオ・ソース・ファイルの開始時間に関するそのセグメントの開始時間をフィールド３３０において表示する。更に、ス

ピーカ・ターン・データベース３００は、フィールド３４０において各セグメントと関連するスピーカを、フィールド３５０における対応するスピーカ・スコアと共に識別する。１つのインプリメンテーションでは、スピーカ・ターン・データベース３００はフィールド３６０において各セグメントと関連する１つ又は複数の代替えスピーカ（次の最適な予測）を、フィールド３７０における対応する代替えスピーカ・スコアと共に識別する。

【００２１】Ａ．スピーカ登録プロセス

図４はスピーカを登録又はエンロールするために使用される既知のプロセスを示す。図４に示されるように、各登録されたスピーカに対して、スピーカの名前が、パルス・コード変調（ＰＣＭ）ファイルのようなスピーカ・トレーニング・ファイルと共にスピーカ登録プロセス４１０に供給される。スピーカ登録プロセス４１０はスピーカ・トレーニング・ファイルを分析し、スピーカ・データベース４２０において各スピーカに対するエントリを作成する。スピーカの音声サンプルをスピーカ・データベース４２０に加えるプロセスは登録と呼ばれる。その登録プロセスはオフラインであり、スピーカ識別システムは、関連するすべてのスピーカに対してそのようなデータベースが存在するものと仮定する。わずかな価値のオーディオに関して、一般には、各スピーカが複数のチャネル及びマイクロフォンから複数の音響的条件を包含することを要求される。登録されたスピーカのトレーニング・データ又はデータベースは、それらのモデルへのアクセスが効率的な認識及び検索のために最適化されるように階層構造を使用して記憶される。

【００２２】Ｂ．プロセス

前述のように、図５に示された同時転写、セグメンテーション及びスピーカ識別プロセス５００は、転写エンジン５１５，セグメンテーション・プロセス６００（図６）及びスピーカ識別プロセス７００（図７）の実行を調整する。同時転写、セグメンテーション及びスピーカ識別プロセス５００は、オーディオ・コーパス・データベース１５０における１つ又は複数のオーディオ・ファイルを分析し、各セグメントと関連するスピーカを表すオーディオ情報の転写をリアルタイムで作成する。図５に示されるように、同時転写、セグメンテーション及びスピーカ識別プロセス５００は、先ず、ステップ５１０においてオーディオ・ファイルからセプストラル（ｃｅｐｓｔｒａｌ）フィーチャを既知の方法で抽出する。一般に、ステップ５１０はオーディオ信号のドメインを時間的ドメインから周波数ドメインに変更し、種々の周波数帯における信号エネルギを分析し、その信号のドメインをセプストラル・ドメインに変更するためにもう１つの変換を使用する。

【００２３】図５に示されるように、ステップ５１０は、転写エンジン５１５，セグメンテーション・プロセス６００（図６）及びスピーカ識別プロセス７００（図７）に共通のフロント・エンド処理を提供する。一般に、ステップ５１０において計算されたフィーチャ・ベクトルは、転写エンジン５１５，セグメンテーション・プロセス（図６）及びスピーカ識別プロセス７００（図７）に対応する３つの多重処理スレッドに分配可能である。それらのフィーチャ・ベクトルは、例えば、計算されたフィーチャ・ベクトルを各チャネル（各処理スレッドに対応する）に分配するためにサーバのような態様で作用する共用メモリ・アーキテクチャを使用して３つの多重処理スレッドに分配可能である。

【００２４】ステップ５１０において生成されたフィーチャ・ベクトルは、多重スレッド環境において並列ブランチに沿って処理される。図５に示され且つ後述されるように、生成されたフィーチャ・ベクトルは多重スレッドを使用して

（ｉ）ステップ５１５において転写エンジン、（ii）ステップ５３０において、図６に関連して後述されるスピーカ・セグメンテーション・プロセス６００、及び（iii）ステップ５６０において、図７に関連して後述されるスピーカ識別プロセス７００に適用される。

【００２５】ステップ５１５において、それらの生成されたフィーチャ・ベクトルは、タイム・スタンプされたワードの転写ファイルを生成するために、ＩＢＭ社から商業的に入手可能なＶｉａＶｏｉｃｅ（商標）音声認識システムのような転写エンジンに供給される。しかる後、それらのタイム・スタンプされたワードは、ステップ５２０においてタイム・スタンプ・ワード・データベース２００の中に任意選択的に収集可能である。更に、それらのタイム・スタンプされたワードは後述のステップ５４０においてインタリーバに供給される。

【００２６】生成されたフィーチャ・ベクトルは、ステップ５３０において、図６に関連して後述されるセグメンテーション・プロセス６００に適用される。一般に、セグメンテーション・プロセス６００はスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別する。セグメント境界が存在する各フレームはターンと呼ばれ、同種の各セグメントは単一のスピーカの音声に対応しなければならない。一旦セグメンテーション・プロセス６００によって描出されると、各セグメントは（そのセグメントがスピーカ認識システムに対して要求される最小セグメント長の要件に合致すると仮定して）特定のスピーカによって発声されたものとして分類可能である。

【００２７】セグメンテーション・プロセス６００によって識別されたターンは、登録されたスピーカ・データベース４２０を使用して各セグメントにスピーカ・ラベルを割り当てるために、ステップ５１０において生成されたフィーチャ・ベクトルと共に、ステップ５６０において、図７と関連して後述されるスピーカ識別プロセス７００に適用される。一般に、スピーカ識別システムは

セグメント発声音をスピーカ・データベース４２０に比較し（図４）、「最も近似した」スピーカを見つける。スピーカ識別プロセス７００によって作成されたその割り当てられたスピーカ・ラベルは後述のステップ５５０に供給される。

【００２８】ステップ５１５において転写エンジンによって作成されたタイム・スタンプ済みワードは、ステップ５３０においてセグメンテーション・プロセス６００によって識別されたスピーカ・ターンと共に、ステップ５４０においてインタリーバに適用され、それらのターンをタイム・スタンプ済みワードとインタリーブさせ、切り離された音声セグメントを作成させる。しかる後、切り離された音声セグメント及びステップ５６０においてスピーカ識別システムにより生成されたスピーカ識別子がステップ５５０においてユーザに表示される。

【００２９】１つのインプリメンテーションでは、切り離された音声セグメントは、それらがステップ５４０においてインタリーバによって作成された時にリアルタイムで表示される。更に、例示的な実施例では、そのスピーカ認識システムに対して要求される最小セグメント長は８秒である。従って、一般には、分離された音声セグメントの始まりが最初に与えられた後約８秒で、スピーカ識別ラベルがその転写されたテキストに付加される。切り離された音声セグメントがそのスピーカ認識システムに対して要求される最小セグメント長よりも短い場合、「未定（ｉｎｃｏｎｃｌｕｓｉｖｅ）」のようなスピーカ・ラベルがそのセグメントに割り当て可能であることに注意すべきである。

【００３０】Ｃ．ベイズ情報基準（ＢＩＣ）の背景

前述のように、図６に示されたセグメンテーション・プロセス６００はスピーカを個別化し、非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別する。セグメント境界が存在する各フレームはターンと呼ばれ、同種の各セグメントは単一のスピーカの音声に対応しなければならない。一旦セグメンテーション・プロセス６００によって描出されると、各セグメントは（そのセグメントがスピーカ認識システムに対して要求された最小セグメント長の要件に合致していると仮定して）特定のスピーカによって発声されたものとして分類可能である。セグメンテーション・プロセス６００はベイズ情報基準（ＢＩＣ）のモデル選択基準に基づくものである。ＢＩＣは、ｐ個のパラメータ・モデルのうちのどれがｎ個のデータ・サンプル $x_1,\ldots,x_n$, $x_i \in R^d$ を最もよく表すかを決定するするために使用される漸近的に最適なベイズのモデル選択基準である。各モデルM_iは複数のパラメータk_iを有する。サンプルx_iは独立したものである仮定する。

【００３１】ＢＩＣの原理に関する詳細な検討のためには、例えば、The Annals of Statistics 誌の第６巻４６１－４６４ページ（１９７８）における G.Schwarz 氏による「モデルの寸法の見積り（Estimating the Dimension of a Model）」と題した論文を参照してほしい。そのＢＩＣの原理によれば、十分に大きいｎに対して、データの最良のモデルは次式を最大化するものである。

$BIC_i = \log L_i(x_1,\ldots,x_n) - (\lambda k_i \log n)/2$

但し、$\lambda = 1$であり、L_iはモデルM_iにおけるデータの最大見込み値（換言すれば、M_iのk_iパラメータに対する最大の見込み値をもったデータの見込み値）である。２つのモデルしか存在しない時、モデル選択のために簡単なテストが使用される。特に、$\Delta BIC = BIC_1 - BIC_2$が正である場合、モデル$M_1$がモデル$M_2$に優先して選択される。同様に、$\Delta BIC = BIC_1 - BIC_2$が負である場合、モデル$M_2$がモデル$M_1$に優先して選択される。

【００３２】Ｄ．スピーカ・セグメンテーション

図６に示されたセグメンテーション・プロセス６００は、セグメント境界が存在するすべての可能なフレームを識別する。汎用性を損なうことなく、精々１つのセグメント境界しか存在しない連続したデータ・サンプル$(x_1,\ldots,x_n)$のウインドウを考察する。

【００３３】フレームｉにおいてセグメント境界が存在するかどうかに関する基本的な疑問が次のような２つのモデル、即ち、モデルM_1及びモデルM_2の間のモデル選択問題として生じ得る。なお、モデルM_1は$(x_1,\ldots,x_n)$が単一の全共分散ガウス分布から得られる場合であり、モデルM_2は$(x_1,\ldots,x_i)$が第１ガウス分布から得られ、$(x_{i+1},\ldots,x_n)$が第２ガウス分布から得られることによって$(x_1,\ldots,x_n)$が２つの全共分散ガウス分布から得られる。

【００３４】$x_i \in R^d$であるので、モデルM_1は$k_1 = d + d(d+1)/2$のパラメータを有し、一方、モデルM_2は２倍のパラメータ（$k_2 = 2k_1$）を有する。次式が負である場合、ｉ番目のフレームがセグメント境界に対する良好な候補であることがわかる。

【数１】

$$\Delta BIC_i = -\frac{n}{2}\log|\Sigma_w| + \frac{i}{2}\log|\Sigma_f| + \frac{n-i}{2}\log|\Sigma_s|$$

$$+\frac{1}{2}\lambda\left(d+\frac{d(d+1)}{2}\right)\log n$$

【００３５】但し、$|\Sigma_w|$はウインドウ全体（即ち、ｎ個のフレームすべて）の共分散の行列式である。$|\Sigma_f|$はそのウインドウの第１サブディビジョンの共分散の行列式であり、$|\Sigma_s|$はそのウインドウの第２サブディビジョンの共分散の行列式である。

【００３６】従って、ステップ６１０において、２つのサブサンプル（$x_1,\ldots,x_i$）及び（$x_{i+1},\ldots,x_n$）が連続的なデータ・サンプル（$x_1,\ldots,x_n$）のウインドウから設定される。セグメンテーション・プロセス６００はステップ６１５乃至６２８において数多くのテストを行い、境界の検出があまりありそうもないロケーションにそのウインドウにおけるいくつかのＢＩＣテストが対応する時、それらのテストを排除する。特に、ステップ６１５において、可変数αの値が（ｎ／ｒ）－１の値に初期設定される。但し、ｒは（フレームにおける）検出解像度である。しかる後、ステップ６２０において、その値αが最大値α_{max}を越えるかどうかを決定するためのテストが行われる。ステップ６２０において値αが最大値α_{max}を越えることが決定される場合、ステップ６２４において、カウンタｉが$(\alpha-\alpha_{max}+1)$ｒの値に設定される。しかし、ステップ６２０において、値αが最大値α_{max}を越えないことが決定される場合、ステップ６２８において、カウンタｉがｒの値に設定される。しかる後、ステップ６３０において、上記の式を使用してＢＩＣ値における差が計算される。

【００３７】ステップ６４０において、カウンタｉの値がｎ－ｒに等しいかどうか、換言すれば、ウインドウにおけるすべての可能なサンプルが評価されてしまったかどうかを決定するためのテストが行われる。ステップ６４０においてカウンタｉの値がｎ－ｒに等しくないことが決定される場合、ステップ６５０においてそのｉの値がｒだけインクレメントされ、ステップ６３０においてウインドウにおける次のサンプルに対する処理を継続する。しかし、ステップ６４０においてカウンタｉの値がｎ－ｒに等しいことが決定される場合、ステップ６６０において、ＢＩＣ値における最小の差（ΔBIC_{i_0}）が負であるかどうかを決定するための更なるテストが行われる。ステップ６６０において、ＢＩＣ値における最小の差（ΔBIC_{i_0}）が負でないことが決定される場合、新しいウインドウを上記方法で考察するためにステップ６１０へ戻る前に、ステップ６６５においてウインドウ・サイズが増加させられる。従って、１つのウインドウにおけるすべてのカウンタｉに対するΔＢＩＣ 値が計算され、それらのうちのいずれも負のΔＢＩＣ 値をもたらすものでない時、ウインドウ・サイズｎが増加させられるだけである。

【００３８】しかし、ステップ６６０において、ＢＩＣ値における最小の差が負であることが決定される場合、ステップ６７０において、i_0がセグメント境界として選択される。しかる後、ステップ６７５において、新しいウインドウの始まりがi_0+1に移り、ウインドウ・サイズがN_0に設定され、その後、新しいウインドウを上記の方法で考察するためにプログラム制御はステップ６１０に戻る。

【００３９】従って、ｉのすべての可能な値に対してＢＩＣ差のテストが行われ、最大の負のΔBIC_iによってi_0が選択される。そのウインドウではフレームｉにおいてセグメント境界が検出可能である。$\Delta BIC_{i_0}<0$である場合、x_{i_0}がセグメント境界に対応する。そのテスト結果が否定的である場合、後述のように、ステップ６６０において更なるデータ・サンプルが（パラメータｎを増加させることによって）現ウインドウに加えられ、フィーチャ・ベクトルがすべてセグメント化されてしまうまで、プロセスはデータ・サンプルのこの新しいウインドウに関して反復される。一般に、ウインドウ・サイズは、自身が１つのウインドウ拡張から別のウインドウ拡張に増加する複数のフィーチャ・ベクトルによって拡張される。しかし、ウインドウは、或る最大値よりも大きい多数のフィーチャ・ベクトルによっては拡張されることはない。ステップ６７０においてセグメント境界が検出された時、ウインドウ拡張値はそれの最小値（N_0）を検索する。

【００４０】Ｅ．可変ウインドウ方式

本発明のもう１つの特徴によれば、特に小さいセグメントにおける全体の精度を改良する新しいウインドウ選択方式が提供される。セグメンテーション・プロセス６００が遂行されるウインドウ・サイズの選択は非常に重要である。その選択されたウインドウがあまりにも多くのベクトルを含む場合、いくつかの境界が脱落することがある。一方、選択されたウインドウが小さ過ぎる場合、情報の不足の結果、ガウス分布によるデータの表示が不十分になるであろう。

【００４１】セグメント境界が検出されなかった場合、一定量のデータを現ウインドウに加えることが提案され

た。そのような方式は、精度を改良するために「前後関係（ｃｏｎｔｅｘｔｕａｌ　ｉｎｆｏｒｍａｔｉｏｎ）」を利用するものではない。セグメント境界が検出されても又はされなくても、或いは境界が長い間検出されなくても、同じ量のデータが加えられる。

【００４２】本発明の改良されたセグメンテーション・プロセスは、新しい境界が生じそうなエリアにおける比較的少量のデータを考察し、新しい境界が生じそうもない時にはウインドウ・サイズをもっと大きく増加させる。先ず、小さいサイズのベクトルのウインドウ（一般には、１００フレームの音声）を考察する。現ウインドウにおいてセグメント境界が検出されない場合、ウインドウのサイズはΔN_iフレームだけ増加する。この新しいウインドウにおいて境界が検出されない場合、フレームの数はΔN_{i+1}だけ増加する。なお、セグメント境界が検出されるまで、又はウインドウ拡張が最大サイズに達してしまうまで（境界が生じる場合に精度の問題を回避するために）、$\Delta N_i=\Delta N_{i-1}+\delta_i$である。但し、$\delta=2\delta_{i-1}$である。これは、ウインドウが依然として小さい時にはかなり遅いウインドウ・サイズの増加及びウインドウが大きくなる時には速いウインドウ・サイズの増加を保証する。ウインドウ内でセグメント境界が検出される時、最小のウインドウ・サイズを使用して次のウインドウがその検出された境界の後に始まる。

【００４３】Ｆ．ＢＩＣテストの効率の改良

本発明のもう１つの特徴によれば、ＢＩＣテストが行われるロケーションの良好な選択によって処理時間全体の改良が得られる。ウインドウにおけるＢＩＣテストのうちの或るものは、境界の検出がありそうもないロケーションにそれらが対応する時、任意に排除可能である。先ず、ＢＩＣテストは各ウインドウの境界においては行われない。それは、それが非常にわずかなデータでもって１つのガウス分布を必ず表示するためである（この明らかに小さいゲインがセグメント検出を通して繰り返され、実際には、それは無視し得るパフォーマンス・インパクトを持たない）。

【００４４】更に、現ウインドウが大きい時にＢＩＣテストがすべて行われる場合、何らかの新しい情報が加えられる度に、そのウインドウの開始時においてＢＩＣ計算が何回も行われたであろう。例えば、１０秒のウインドウ・サイズにおいて最初の５秒内にセグメント境界が検出されなかった場合、１０秒の現ウインドウの拡張によって、最初の５秒内に境界が認められるということは全くありそうもない。従って、（ウインドウ拡張に続く）現ウインドウの始まりにおけるＢＩＣ計算を無視することによってＢＩＣ計算の数を減少させることができる。実際には、ＢＩＣ計算の最大数は、今や、必要とされる速度／精度レベルに従って調整された調節可能なパラメータ（図３におけるα_{max}）である。

【００４５】従って、セグメンテーション・プロセス６００は、セグメンテーション情報に関する何らかのフィードバックを持つ前にそれが必要とする最大時間を知ることを可能にする。それは、たとえセグメント境界が検出されなくても、ウインドウが十分に大きい場合、第１フレームにセグメントが存在しないということがわかるためである。この情報は速度信号のうちのこの部分に関して別の処理を行うために使用可能である。

【００４６】Ｇ．ＢＩＣペナルティ・ウェート

ＢＩＣの式は、理論と基準に関する実用的な応用との間の差を補うために、ペナルティ・ウェート・パラメータλを利用する。ミス率と誤警報率との間の良好なトレード・オフを与えるλの最良値は１．３であることがわかっている。放送ニュースの転写に対するセグメンテーション精度に関するλの影響をより総合的に研究するためには、M.S.Thesis, Institut Eurecom 誌（フランス、１９９８）における A. Tritschler 氏による「ＢＩＣを使用したセグメンテーション・イネーブルド音声認識アプリケーション（A Segmentation-Enabled Speech Recognition Application）」と題した論文を参照してほしい。

【００４７】原則として、係数λはタスク依存のものであり、新しいタスク毎に戻されなければならないけれども、実際には、そのアルゴリズムは種々のタイプのデータに適用されており、同じ値のλを使用することによるパフォーマンスにおける認め得る程度の変化は存在しない。

【００４８】Ｈ．スピーカ識別プロセス

前述のように、同時転写、セグメンテーション及びスピーカ識別プロセス５００は、ステップ５６０において、図７に示されたスピーカ識別プロセス７００を実行し、登録されたスピーカ・データベース４２０を使用して各セグメントにスピーカ・ラベルを割り当てる。図７に示されるように、スピーカ識別プロセス７００は、ステップ５１０において共通のフロント・エンド・プロセッサによって生成されたフィーチャ・ベクトルと共に、セグメンテーション・プロセス６００によって識別されたターンを受け取る。一般に、スピーカ識別システムはスピーカ・データベース４２０（図４）にセグメント発声音を比較し、「最も近似した」スピーカを検出する。

【００４９】ターン及びフィーチャ・ベクトルは、ステップ７１０において、単一のスピーカによる音声のチャンクより成るセグメント発声音を形成するように処理される。ステップ７２０において、セグメント発声音がスピーカ識別システムに供給される。スピーカ識別システムを検討するためには、例えば、Proc. of Speaker Recognition and Its Commercial and Forensic Application, Avignon, France (1998) 誌における H.S.M.Beigi 氏他による「ＩＢＭモデル・ベース及びフレーム毎のスピーカ認識（IBM Model-Based and Frame-By-Frame Speaker-Recognition）」と題した論文を参照してほしい。

一般に、スピーカ識別システムはセグメント発声をスピーカ・データベース４２０（図４）に比較し、「最も近似した」スピーカを検出する。

【００５０】スピーカ識別システムは２つの異なるインプリメンテーション、即ち、モデル・ベース方式及びフレーム・ベース方式を有し、それらは付随した利点及び欠点を有する。エンジンは、放送ニュースのような番組の生のオーディオ・インデキシングを容易にするために独立したテキスト及び言語の両方である。

【００５１】Ｉ．スピーカ識別（モデル・ベース方式）
スピーカの母集団に対して一組のトレーニング・モデルを作成するために、下記のようなｄ次元のフィーチャ・ベクトルを持ったＭ個の音声フレームのシーケンスに基づいたｉ番目のスピーカに対するモデルM_iが計算される。

【数２】

$$\left\{\vec{f}_m\right\}_{m=1,\ldots,M}$$

【００５２】これらのモデルは、ガウス分布が選択される場合に対して、平均ベクトル、共分散マトリックス、及びカウントより成る下記のようなそれらの統計的パラメータによって記憶される。なお、各スピーカｉはn_i個の分布よりなるモデルでもって終わり得るものである。

【数３】

$$\left\{\vec{\mu}_{i,j}, \Sigma_{i,j}, \vec{C}_{i,j}\right\}_{j=1,\ldots,n_i}$$

【００５３】２つのそのようなモデルを比較するために、Proc. ICASSP98 誌（Seattle, WA, 1998）における H.S.M. Beigi 氏他による「分布の集合体相互間の距離測定法及びスピーカ認識に対するそれの応用（A Distance Measure Between Collections of Distributions and Its Application to Speaker Recognition）」と題した論文において提案された距離測定法を使用して、スピーカ識別（クレームを実証する）、スピーカ分類（スピーカを割り当てる）、スピーカ検証（ラベルされたスピーカの特性に匹敵する特性を持ったスピーカの「コーホート（ｃｏｈｏｒｔ）」セットとラベルを比較することによって分類を確認するための第２パス）、及びスピーカ・クラスタリングを含む多くの種々な機能を持ったスピーカ認識システムを考案するために階層構造が作成される。

【００５４】スピーカ認識のために考案されたその距離測定法は、異なる数の分布n_iを持った受容可能な距離の計算を可能にする。２つのスピーカをそれらのパラメータ的な表示に基づいて比較するだけで、２つのスピーカを比較するというそのタスクを計算主体でないものにするという特徴を常に持つ必要がなくなる。しかし、認識段階に対するこの距離測定法の欠点は、比較の計算が始まる前に個々のテストのモデル（要求者：Ｃｌａｉｍａｎｔ）を形成するために音声セグメント全体が使用されなければならないということである。フレーム・ベース方式はこの欠点を緩和する。

【００５５】Ｊ．スピーカ識別（フレーム・ベース方式）
M_iをｉ番目の登録されたスピーカに対応するモデルであると仮定する。M_iは、スピーカｉのガウス混合モデル（ＧＭＭ）のn_i個のコンポーネントの各々に対する平均ベクトル、共分散マトリクス、及び混合ウェートより成る次のようなパラメータ・セットによって全体的に定義される。

【数４】

$$\left\{\vec{\mu}_{i,j}, \Sigma_{i,j}, \vec{p}_{i,j}\right\}_{j=1,\ldots,n_i}$$

【００５６】これらのモデルは、前のセクションにおいて説明したように、下記のようなｄ次元のフィーチャ・ベクトルを持ったＭ個の音声フレームのシーケンスより成るトレーニング・データを使用して作成される。

【数５】

$$\left\{\vec{f}_m\right\}_{m=1,\ldots,M}$$

【００５７】スピーカ母集団のサイズがN_sである場合、モデル・ユニバースのセットは次のようになる。

【数６】

$$\{M_i\}\ \ i=1,\ldots,Np$$

【００５８】基本的な目的は、次式のようなＮ個のフレームのシーケンスとして表されたテスト・データをM_iが最もよく示しているというようなｉを見つけること、及びそれらのモデルのうちデータを十分に記述するものがないという決定を行うことである。

【数７】

$$\left\{\vec{f}_n\right\}_{n=1,\ldots,N}$$

【００５９】次のようなフレーム・ベースのウェート付けされた距離測定法$d_{1,n}$はその決定を行う場合に使用される。

【数８】

$$d_{i,n} = -\log\left[\sum_{j=1}^{n_i} p_{i,j} p\left(f_n \middle| j^{th}\ component\ of\ \vec{M}_i\right)\right]$$

【００６０】但し、正規の表示を使用すると、次のようになる。

【数９】

$$p\left(\vec{f}_n \middle| \cdot\right) = \frac{1}{(2\pi)^{d/2}\left|\sum_{i,j}\right|^{1/2}} e^{-\frac{1}{2}\left(\vec{f}_n - \vec{\mu}_{i,j}\right)^t \sum_{i,j}^{-1}\left(\vec{f}_n - \vec{\mu}_{i,j}\right)}$$

【００６１】テスト・データからのモデルM_iの合計距離D_iはテスト・フレームの合計数を超えたすべての距離の和であると見なされる。

【００６２】分類のために、音声セグメントのモデルまでの最小距離を持ったモデルが選択される。その最小距離を背景モデルの距離に比較することによって、オリジナル・モードのうちのいずれも十分に合致しないことを表示するための方法を提供することが可能である。別の方法として、合計距離を計算するために投票集計技法が使用可能である。

【００６３】検証のために、ラベルされたスピーカのコーホートを形成する所定セットのメンバが種々のバックグラウンド・モデルでによって増大する。このセットをモデル・ユニバースとして使用すると、テスト・データは、要求者（Ｃｌａｉｍａｎｔのモデルが最小距離を有するかどうかをテストすることによって検証される。そうでない場合、それは拒絶される。

【００６４】この距離測定法は、スピーカ相互間の距離を計算するために音声のフレームが保持されなければならないので、トレーニングでは使用されない。従って、トレーニングは、前述のモデル・ベースのテクニックのための方法を使用して行われる。

【００６５】ステップ７２０において生成されたその割り当てられたスピーカ・ラベルは、下記のように、ユーザへ出力するために任意選択的にブロック５５０（図５）への暫定的提供が可能である。ステップ７３０において、その割り当てられたスピーカ・ラベルは、スピーカ分類の結果に関して第２パスを行うことによって検証される。ステップ７３０においてスピーカ識別が検証される場合、そのスピーカ・ラベルはユーザへの出力のためにブロック５５０（図５）に供給される。更に、ステップ７４０において、オリジナルの登録されたスピーカ・モデルからオーディオ・テスト・セグメントまでの距離を表す割り当てられたスコアと共に、最善の選択を表すエントリ、又は、望ましい場合には、代替えの選択を表すエントリを、スピーカ・ターン・データベース３００において任意選択的に作成することが可能である。

【００６６】本願において開示され及び図示された実施例並びにその変形は単に本発明の原理を説明するものであること、及び本発明の技術的範囲及び精神から逸脱することなく種々の修正を当業者が実施することが可能であることは理解されるべきである。

【００６７】まとめとして、本発明の構成に関して以下の事項を開示する。

【００６８】（１）１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、を含む方法。

（２）前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写を作成することを特徴とする請求項１に記載の方法。

（３）前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項１に記載の方法。

（４）前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項１に記載の方法。

（５）前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項１に記載の方法。

（６）前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも１つの別の選択を割り当てるステップを更に含むことを特徴とする請求項１に記載の方法。

（７）前記転写するステップ、識別するステップ、及び割り当てるステップはマルチ・スレッド環境では並列処理ブランチにおいて行われることを特徴とする請求項１

に記載の方法。
（８）前記識別するステップはＢＩＣモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項１に記載の方法。
（９）１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、
（ａ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、
（ｂ）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び（ｃ）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するステップと、を含む方法。
（１０）前記フィーチャ・ベクトルは共用メモリ・アーキテクチャを使用して前記並列処理ブランチに適用されることを特徴とする請求項９に記載の方法。
（１１）前記共用メモリ・アーキテクチャは前記計算されたフィーチャ・ベクトルを前記並列処理ブランチの各々に対応するチャネルに分配することを特徴とする請求項１０に記載の方法。
（１２）前記転写するステップは前記オーディオ・ソースにおける各ワードに対してタイム・アライメントを持った転写物を作成することを特徴とする請求項９に記載の方法。
（１３）前記識別するステップは非同種の音声部分相互間にセグメント境界が存在するすべての可能なフレームを識別することを特徴とする請求項９に記載の方法。
（１４）前記割り当てるステップは登録されたスピーカ・データベースを利用してスピーカ・ラベルを各識別されたセグメントに割り当てることを特徴とする請求項９に記載の方法。
（１５）前記割り当てるステップは前記割り当てられたスピーカ・ラベルの信頼性を表すスコアを割り当てるステップを更に含むことを特徴とする請求項９に記載の方法。
（１６）前記割り当てるステップは前記割り当てられたスピーカ・ラベルに対して少なくとも１つの別の選択を割り当てるステップを更に含むことを特徴とする請求項９に記載の方法。
（１７）前記識別するステップはＢＩＣモデル選択基準を使用してセグメント境界を識別することを特徴とする請求項９に記載の方法。
（１８）１つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、コンピュータ読み取り可能なコードを記憶するメモリと、前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、を含み、前記コンピュータ読み取り可能なコードは、前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するように、前記転写と実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するように、及び各識別されたセグメントにスピーカ・ラベルを割り当てるように、構成されることを特徴とするシステム。
（１９）コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、前記コンピュータ読み取り可能なプログラム・コード手段は、オーディオ情報のテキスト・バージョンを作成するためにオーディオ・ソースを転写するステップと、前記転写するステップと実質的に同時に前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、を含むことを特徴とする製造物。
（２０）１つ又は複数のオーディオ・ソースからオーディオ情報を転写するためのシステムにして、コンピュータ読み取り可能なコードを記憶するメモリと、前記メモリに動作関係に結合され、前記コンピュータ読み取り可能なコードをインプリメントするように構成されたプロセッサと、を含み、前記コンピュータ読み取り可能なコードは、前記オーディオ情報からフィーチャ・ベクトルを計算し、（ｉ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、（ii）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び（iii）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するように構成されることを特徴とするシステム。
（２１）コンピュータ読み取り可能なプログラム・コード手段を組み込まれたコンピュータ読み取り可能な媒体を含み、前記コンピュータ読み取り可能なプログラム・コード手段は、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、（ｉ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、（ii）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び（iii）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するステップと、を含むことを特徴とする製造物。
（２２）１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報のテキスト・バージョンを作成するために前記オーディオ・ソースを転写するステップと、前記オーディオ・ソースにおける潜在的なセグメント境界を識別するステップと、識別された各セグメントにスピーカ・ラベルを割り当てるステップと、前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時

に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、を含む方法。

（２３）１つ又は複数のオーディオ・ソースからオーディオ情報を転写するための方法にして、前記オーディオ情報からフィーチャ・ベクトルを計算するステップと、（ｉ）前記オーディオ・ソースを転写して前記オーディオ情報のテキスト・バージョンを作成するために、（ｉｉ）前記オーディオ・ソースにおける潜在的なセグメント境界を識別するために、及び（ｉｉｉ）各識別されたセグメントにスピーカ・ラベルを割り当てるために、前記フィーチャ・ベクトルを３つの並列処理ブランチに適用するステップと、前記転写するステップ、識別するステップ、及び割り当てるステップと実質的に同時に前記テキスト・バージョンを前記割り当てられたスピーカ・ラベルと共に供給するステップと、を含む方法。

【図面の簡単な説明】

【図１】本発明によるオーディオ転写及びスピーカ分類＊システムのブロック図である。

【図２】図１のタイム・スタンプされたワード・データベースからのテーブルである。

【図３】図１のスピーカ・ターン・データベースからのテーブルである。

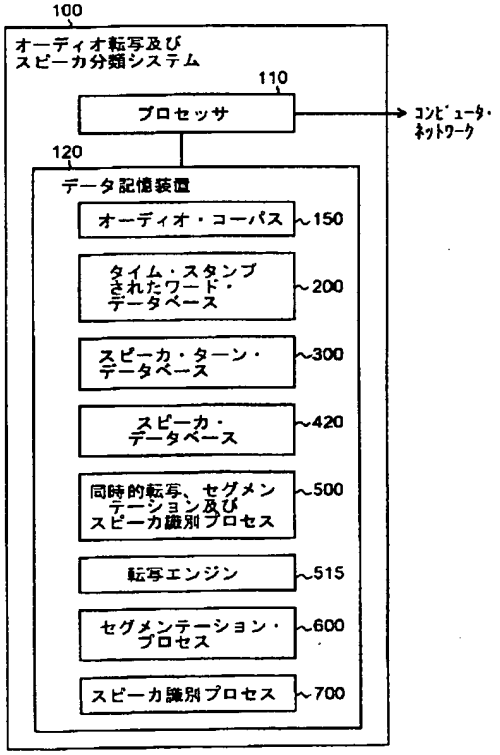
【図４】本発明による代表的なスピーカ登録プロセスを示す。

【図５】図１のオーディオ転写及びスピーカ分類システムによって遂行される例示的な同時転写、セグメンテーション及びスピーカ識別プロセスを説明するフローチャートである。

【図６】図１のオーディオ転写及びスピーカ分類システムによって遂行される例示的なセグメンテーション・プロセスを説明するフローチャートである。

【図７】図１のオーディオ転写及びスピーカ分類システムによって遂行される例示的なスピーカ識別プロセスを説明するフローチャートである。

【図１】



【図２】

200 タイム・スタンプされたワード・データベース

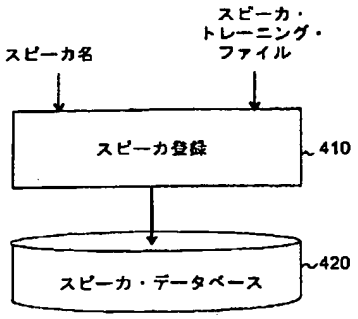
	ワード・ストリング 220	開始時間 230	ワード期間 240
211	1	t_1	
212	2	t_2	
213	⋮	⋮	⋮
214	N	t_N	

【図３】

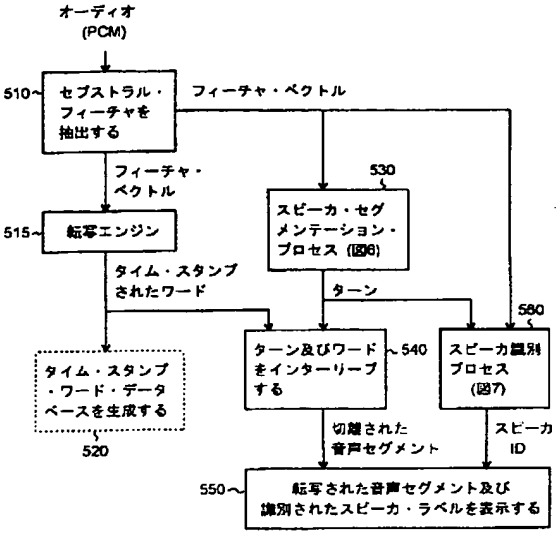
300 スピーカ・ターン・データベース

	セグメント番号 320	開始時間 330	スピーカ・ラベル（最良の推量） 340	スコア 350	別のスピーカ・ラベル 360	別のスピーカ・スコア 370
305	1	T_A	スピーカ X	S_{10}	スピーカ K	$S_{10\text{-}a}$
306	2	T_K	スピーカ K	S_{11}	スピーカ L	$S_{11\text{-}a}$
307	⋮	⋮	⋮	⋮	⋮	⋮
308	N	T_E	スピーカ G	S_{12}	スピーカ P	$S_{12\text{-}a}$

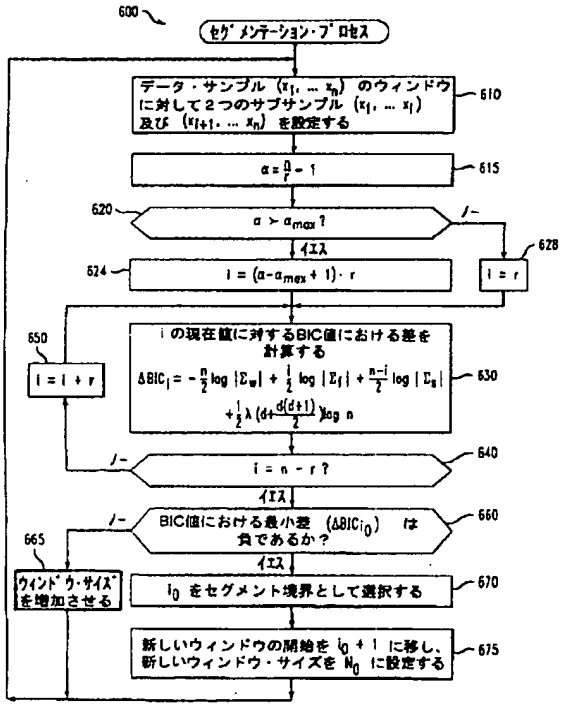
【図４】


スピーカ名
スピーカ・トレーニング・ファイル
スピーカ登録
410
スピーカ・データベース
420

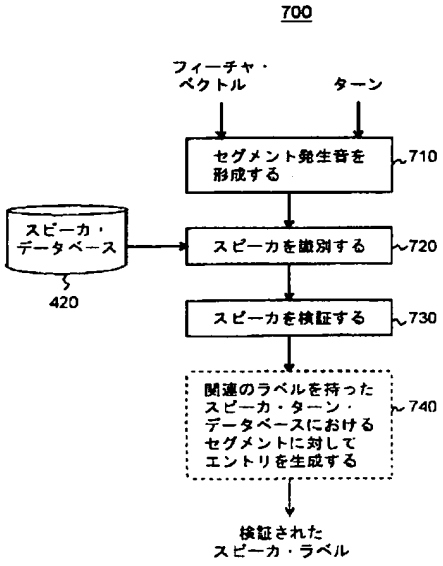
【図５】


オーディオ (PCM)
510
セプストラル・フィーチャを抽出する
フィーチャ・ベクトル
530
スピーカ・セグメンテーション・プロセス (図6)
515
転写エンジン
フィーチャ・ベクトル
タイム・スタンプされたワード
ターン
560
ターン及びワードをインターリーブする
540
スピーカ識別プロセス (図7)
タイム・スタンプ・ワード・データベースを生成する
520
切離された音声セグメント
スピーカ ID
550
転写された音声セグメント及び識別されたスピーカ・ラベルを表示する

【図６】


600
セグメンテーション・プロセス
データ・サンプル ($x_1, \ldots x_n$) のウィンドウに対して２つのサブサンプル ($x_1, \ldots x_i$) 及び ($x_{i+1}, \ldots x_n$) を設定する
610
$\alpha = \frac{n}{r} - 1$
615
620
$\alpha > \alpha_{max}$?
ノー
イエス
628
$i = r$
624
$i = (\alpha - \alpha_{max} + 1) \cdot r$
650
$i = i + r$
i の現在値に対するBIC値における差を計算する
$\Delta BIC_i = -\frac{n}{2}\log|\Sigma_w| + \frac{i}{2}\log|\Sigma_1| + \frac{n-i}{2}\log|\Sigma_2| + \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log n$
630
$i = n - r$?
640
イエス
ノー
BIC値における最小差 (ΔBIC_{i_0}) は負であるか？
660
665
ウィンドウ・サイズを増加させる
イエス
i_0 をセグメント境界として選択する
670
新しいウィンドウの開始を $i_0 + 1$ に移し、新しいウィンドウ・サイズを N_0 に設定する
675

【図７】


700
フィーチャ・ベクトル
ターン
セグメント発生音を形成する
710
スピーカ・データベース
420
スピーカを識別する
720
スピーカを検証する
730
関連のラベルを持ったスピーカ・ターン・データベースにおけるセグメントに対してエントリを生成する
740
検証されたスピーカ・ラベル

フロントページの続き

(72)発明者 アラン・シャルル・ルイ・トレザー
アメリカ合衆国ニューヨーク州、ニューヨーク、ウエスト・シックスティサード・ストリート、243 ナンバー・５・エイ

(72)発明者 マハシュ・ヴィズワナザン
アメリカ合衆国ニューヨーク州、ヨークタウン・ハイツ、ダグラス・ドライブ 3024